Environmental Health Perspectives Supplements
Vol. 102 Suppl 1:19–23 1994

Notes on the Use of Historical Controls

by Isao Yoshimura[1] and Kazuhiko Matsumoto[2]

The principles and methods of incorporating historical controls in four cases (C1, stable control; C2, rare occurrence of responses; C3, small group size; C4, historical control as a reference) are discussed. Two points are emphasized: one is that the historical control should be regarded as a given condition and the other is that the historical control should be used conservatively. Incorporating historical controls is recommended only when it is advantageous under the conditional evaluation of the performance and even in the conservative use of controls. For case C1, adjusting the critical value for the Cochran-Armitage trend test is proposed. For case C2, a modified conditional trend test proposed by Yanagawa et al. is appreciated as a proper procedure. For case C3, a conservative use of interblock information is discussed. The incorporation of the historical control is not recommended for case C4.

Introduction

Ever since the work of R. A. Fisher, we have been analyzing, in principle, the data in a toxicological experiment independently of other experiments. The number of well-controlled experiments conducted under the same protocol has recently increased, however, and this has tempted us to incorporate the data from past experiments with the data from current experiments. The temptation is especially strong in the following cases, where control groups in past experiments are referred to as the historical control. Incorporating the historical control would serve to *a*) increase the power of hypothesis testing when the historical control is stable, as is the case reported in Hayashi et al. (*1*): C1; *b*) carry out hypothesis testing when the occurrence of response is rare, as is the case in Tarone (*2*) or Yanagawa and Hoel (*3*): C2; *c*) increase the power of hypothesis testing when the group size is small, as is the case when the experiment is conducted on dogs: (C3; *d*) validate the judgment that the observed significance is a realization of type I errors, as is the case when so many items are tested that an inflated type I error is likely to occur: C4. Experiencing many such cases, researchers engaged in toxicological experiments ask statisticians the following questions: In what cases can we use the historical control? How should we incorporate the historical control? The purpose of this paper is to address these questions.

The Meaning of "Historical"

Because statistical testing is a principal concern in the data analysis in the cases mentioned above, we concentrate our attention in this paper on testing procedures. There are many works in the literature addressing the use of the historical control in toxicological experiments. Among them, Margolin and Risko (*4*) presented a good summary of the general principle. For carcinogenicity studies, Tarone (*2*), Yanagawa and Hoel (*3*), Tamura and Young (*5*), Hoel and Yanagawa (*6*), Krewski et al. (*7*), and Yanagawa et al. (*8*) proposed some procedures or discussed problems to be considered based on the β-binomial model.

In their arguments (*2–8*), however, the time dependence of the historical control in comparison to the current experiment is not consciously considered. The arguments, except the one in Yanagawa et al. (*8*), are valid even when we reanalyze the data in a past experiment by incorporating data in succeeding experiments. In real situations, incorporation of the historical control is sought after only when a laboratory has collected enough data from past experiments and has confirmed some sort of homogeneity of past experiments. Once the historical control is saved in a database, researchers in that laboratory always refer to the same data repeatedly in the data analysis of succeeding experiments. In this circumstance, the data in the historical control are not randomly realized values but refer to a given condition fixed in advance in the data analysis of the current experiment. To address the questions mentioned earlier, we should examine the performance of each procedure from the viewpoint that the historical control is a given, fixed condition. This viewpoint of conditional use is the first point of our assertion.

The historical control is used to estimate unknown parameters related to the current experiment. The assertion that the historical control should be regarded as a given, fixed condition means that we should regard the estimated values based on the historical control as including some deviations from true parameter values, though the amount of deviation is within random variations. Because the historical control is a realization of random variables, both negative deviation and positive deviation may occur. In general, when the positive deviation causes the test based on the historical control to be conservative, the negative deviation causes the test to be liberal, and vice versa. Because we

[1] Faculty of Engineering, Science University of Tokyo, 1-3 Kagurazaka, Tokyo 162, Japan.
[2] Research Laboratory, Toyo-Jozo Co. Ltd., Oohito, Shizuoka, 410-23, Japan.
Address reprint requests to I. Yoshimura, Faculty of Engineering, Science University of Tokyo, 1-3 Kagurazaka, Tokyo 162, Japan.
This paper was presented at the International Biostatistics Conference on the Study of Toxicology that was held May 13–25, 1991, in Tokyo, Japan.

cannot know which situation is realized, we have to design a test procedure that is conservative in the sense that I errors are controlled within a target significance level even when a disadvatageous deviation has occurred. This viewpoint of conservative use is the second point of our assertion.

Stable Historical Control

To make the arguments clear, we deal only with simple cases. Assume that the current experiment consists of $(a + 1)$ groups, $A_0, A_1, \ldots A_a$ of n individuals and that each individual in the group A_i is exposed to dose $d_i (d_0 < d_1 < \cdots < d_a$ of a chemical. Let the observed response for the jth individual in A_i be 1 or 0 with probability π_i or $1-\pi_i$, where $\pi_i = \pi(d_i)$, and the responses are independent. We can reduce the observed responses to the random variable $X = (X_0, X_1, \cdots, X_a)$, where X_i is distributed binomially with mean $\mu_i = n\pi_i = \mu(d_i)$. Likewise, assume that the historical control consists of b groups, $B_1, B_2, \cdots B_b$, of controls in b previous experiments. Let the observed variable be $Y = (Y_1, Y_2, \cdots, Y_b)$, where Y_j is distributed binomially with mean $\mu_{(j)} = n\pi_{(j)}$. Assume the X's and Y's are all mutually independent.

Under this formulation, it must be reasonable to regard the case where $\mu_{(1)} = \mu_{(2)} = \cdots = \mu_{(b)} = \mu_0$ as the case C1, that is, the case with the stable historical control. In this case, the observed variable can be reduced to $(\mathbf{X},\mathbf{Y}) = (X_0, X_1, \cdots, X_a, Y)$, where $Y = \Sigma Y_i$ is distributed binominally B (bn, π_0). In most practical situations, when the researcher tries to test the null hypothesis

$$H_0: \mu_0 = \mu_1 = \cdots = \mu_a$$

against an alternative hypothesis

$$H_1: \mu_0 \leq \mu_1 \leq \cdots \leq \mu_a,$$

at least one strict inequality holds, based on $(\mathbf{X},\mathbf{Y})$.

A typical procedure for this problem is the Cochran-Armitage trend test because it is the uniformly most powerful, unbiased test against logistic alternatives (*9*). When the group size, n, is so large that the normal approximation on the binomial distribution is available, two procedures, say Tc and Th, of the trend test with significance level α can be considered by excluding or including Y as follows:

Test To: Reject H_0 if

$$T_C = \frac{\sum_{i=0}^{a} (d_i - d_c)X_i}{\sqrt{\{\sum_{i=0}^{a} (d_i - d_c)^2\} n P_c (1-P_c)}} > u(\alpha) \tag{1}$$

where $d_c = (d_0 + d_1 + \cdots + d_a)/(a+1)$, $P_c = (X_0 + X_1 + \cdots + X_a)/(an+n)$ and $u(\alpha)$ is the upper $100\alpha\%$ point of N(0, 1).

Test Th: Reject H_0 if

$$T_h = \frac{\sum_{i=0}^{a} (d_i - d_h)X_i + (d_0 - d_h)Y}{\sqrt{\{\sum_{i=0}^{a} (d_i - d_h)^2 + b(d_0 - d_h)^2\} n P_h (1-P_h)}} > u(\alpha) \tag{2}$$

where $d_h = \{(b+1)d_0 + d_1 + \cdots + d_a\}/(a+b+1)$ and $P_h = (Y + X_0 + X_1 + \cdots + X_a)/(an+n+n)$.

If Y is regarded as a random variable, the test Th is obviously better than the test Tc. But, if Y is regarded as a given constant y, the type I error of the test Th is not controlled within a target significance level α as is explained below.

Under H_0, the statistic T_h can be written as

$$T_h = \frac{\sum_{i=0}^{a} (d_i - d_h)X_i + (d_0 - d_h)Y}{\sqrt{\{\sum_{i=0}^{a} (d_i - d_h)^2 + b(d_0 - d_h)^2\} n\pi_0 (1-\pi_0)}} + o(1) \tag{3}$$

where o(1) implies a term that tends toward 0 in probability if n tends toward ∞.

Let

$$B^2 = \frac{b(d_0 - d_h)^2}{\sum_{i=0}^{a} (d_i - d_h)^2}, \quad D_X = \frac{\sum_{i=0}^{a} (d_i - d_h)(X_i - n\pi_0)}{\sqrt{\{\sum_{i=0}^{a} (d_i - d_h)^2\} n\pi_0 (1-\pi_0)}} \tag{4}$$

and

$$D_Y = \frac{(Y - bn\pi_0)}{\sqrt{b\ n\pi_0 (1-\pi_0)}} \tag{4a}$$

Then,

$$T_h = \frac{D_X}{\sqrt{1+B^2}} + \frac{D_Y}{\sqrt{1+B^{-2}}} + o(1) \tag{5}$$

B can be regarded as a small quantity, because b is generally much greater than a in the situation where the use of the historical control is asked for and that

$$B^2 = \frac{(a+1)w(1-w)(d_0 - d_c)^2}{\sum_{i=0}^{a} (d_i - d_c)^2 + bw(1-w)(d_0 - d_c)^2} \tag{6}$$

where $w = b/(a+b+1)$. For example, if $a = 4$, $b = 20$ and $d_i = i$, then $B^2 = 0.14$. Under such a situation, T_h is approximated by

$$T_h = D_X + B\ D_Y \tag{7}$$

Because D_x is distributed as N(0,1) independently of D_Y, the conditional distribution of T_h given $Y = y$, is positively (or negatively) biased from N(0,1) if $y > bn\pi_0$ (or $y < bn\pi_0$).

Table 1. Type I error for the tests Tc, Th, and Ta. [a]

	D_y						
Test	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5
Tc	0.052	0.052	0.052	0.052	0.052	0.052	0.052
Th	0.138	0.099	0.066	0.043	0.029	0.018	0.011
$\Phi(BD_y-u)$	0.139	0.102	0.072	0.050	0.033	0.022	0.014
Ta	0.050	0.032	0.020	0.011	0.006	0.004	0.002

[a] $a=4$, $b=20$, $d_i=i$, $\pi_0=0.05$, n=100. The target significance level is 0.05. The replication in the Monte-Carlo simulation is 40,000.

Therefore, type I error given Y = y is approximately evaluated as

$$\Pr\{T_h>u(\alpha)|H_0,Y=y\}= \Pr\{D_X>u(\alpha)-BD_y|H_0,Y=y\}=\Phi\{BD_y-u(\alpha)\} \quad (8)$$

where

$$D_y=\frac{(y-bn\pi_0)}{\sqrt{bn\pi_0(1-\pi_0)}} \quad (9)$$

and Φ is the distribution function of N(0,1) because D_y is distributed as N(0,1), $|D_y|<1.645$ with probability 0.90. For $d_i = i$ and for D_y within this range, some numerical values of the right-hand side of Equation 8 are, together with type I error obtained by a Monte-Carlo simulation, shown in Table 1. Table 1 shows the possibility of an inflation of the type I error.

One idea to control the type I error within the significance level α is to adjust the critical value of the test Th. If we evaluate a possible maximum value of D_y by $u(\alpha')$, we obtain an adjusted test Ta as follows:

Test Ta: Reject H_0 if

$$T_h > u(\alpha)+Bu(\alpha') \quad (10)$$

We think that a reasonable value of α' is 0.05, which gives $u(\alpha')=1.645$. According to the above argument, the incorporation of the historical control is advantageous only when the power of the test Ta is greater than that of the test Tc.

Because, under H_1 and for given y, the statistic T_c is approximately normally distributed with

$$E\{T_c|H_1\}=\frac{\sum_{i=0}^{a}(d_i-d_c)n\pi_i}{\sqrt{\sum_{i=0}^{a}(d_i-d_c)^2n\pi_c(1-\pi_c)}}+o(1) \quad (11)$$

and

$$V\{T_c|H_1\}=\frac{\sum_{i=0}^{a}(d_i-d_c)^2n\pi_i(1-\pi_i)}{\sum_{i=0}^{a}(d_i-d_c)^2n\pi_c(1-\pi_c)}+o(1) \quad (12)$$

the power of the test Tc is given by $\phi(\lambda_c)$, where $\pi_c=\Sigma\pi_i/(a+1)$,

$$\lambda_c=[E\{T_c|H_1\}-u(\alpha)]/\sqrt{V\{T_c|H_1\}} \quad (13)$$

and o(1) is ignored. Similarly, the power of the tests Th and Ta are approximately given by $\Phi(\lambda_h)$ and $\Phi(\lambda_a)$, where

$$E\{T_h|H_1,Y=y\}=\frac{\sum_{i=0}^{a}(d_i-d_h)n\pi_i+(d_0-d_h)y}{\sqrt{\{\sum_{i=0}^{a}(d_i-d_h)^2+b(d_0-d_h)^2\}n\pi_h(1-\pi_h)}}+o(1) \quad (14)$$

$$V\{T_h|H_1,Y=y\}=\frac{\sum_{i=0}^{a}(d_i-d_h)^2n\pi_i(1-\pi_i)}{\{\sum_{i=0}^{a}(d_i-d_h)^2+b(d_0-d_h)^2\}n\pi_h(1-\pi_h)}+o(1) \quad (15)$$

$$\pi_h = w\,\pi_0 + (1-w)\pi_c \quad \text{and} \quad (16)$$

$$\lambda_h=[E\{T_h|H_1,Y=y\}-u(\alpha)]/\sqrt{V\{T_h|H_1,Y=y\}}$$

$$\lambda_a=[E\{T_h|H_1,Y=y\}-u(\alpha)-Bu(\alpha')]/\sqrt{V\{T_h|H_1,Y=y\}} \quad (17)$$

For $d_i = i$, some numerical values of $\Phi(\lambda_c)$, $\phi(\lambda_h)$ and $\Phi(\lambda_a)$ are shown in Table 2, together with values obtained by a Monte-Carlo simulation. Table 2 shows that some parts of increases of the power for the test Th are spurious due to an inflation of the type I error and that the advantage of the incorporation of the historical control is rather limited, even when the historical control is stable enough. According to our assertion, whether the historical control should be incorporated or not is judged through the comparison of the two tests Ta and Tc. The choice is possible

Table 2. Power for the tests Tc, Th, and Ta. [a]

	D_j						
Test	-1.5	-1.0	-.05	0.0	0.5	1.0	1.5
Concave Case[b]							
M-C: Tc	0.390	0.390	0.390	0.390	0.390	0.390	0.390
$\Phi(\lambda_c)$	0.389	0.389	0.389	0.389	0.389	0.389	0.389
M-C: Th	0.965	0.949	0.925	0.897	0.861	0.816	0.767
$\Phi(\lambda_h)$	0.949	0.934	0.913	0.889	0.860	0.837	0.788
M-C: Ta	0.906	0.869	0.823	0.768	0.719	0.650	0.576
$\Phi(\lambda_a)$	0.869	0.839	0.801	0.760	0.714	0.665	0.613
Linear case[c]							
M-C; Tc	0.446	0.446	0.446	0.446	0.446	0.446	0.446
$\Phi(\lambda_c)$	0.442	0.442	0.442	0.442	0.442	0.442	0.442
M-C: Th	0.941	0.915	0.878	0.837	0.790	0.736	0.678
$\Phi(\lambda_h)$	0.918	0.895	0.866	0.834	0.796	0.754	0.707
M-C:Ta	0.851	0.802	0.745	0.681	0.612	0.547	0.476
$\phi(\lambda_a)$	0.809	0.767	0.723	0.674	0.622	0.567	0.510
Convex case[d]							
M-C: Tc	0.468	0.468	0.468	0.468	0.468	0.468	0.468
$\Phi(\lambda_c)$	0.465	0.465	0.465	0.465	0.465	0.465	0.465
M-C: Th	0.882	0.842	0.791	0.734	0.673	0.604	0.540
$\Phi(\lambda_h)$	0.832	0.793	0.750	0.702	0.650	0.594	0.538
M-C: Ta	0.755	0.690	0.617	0.545	0.471	0.404	0.336
$\Phi(\lambda_a)$	0.669	0.617	0.559	0.502	0.444	0.388	0.334

[a] $a=4$, $b=20$, $d_i=i$, $\pi_0=0.05$, n=100. The largest significance level is 0.05. The replication in the Monte-Carlo simulation is 40,000.
[b] $\pi_0=0.0500$; $\pi_1=0.0750$; $\pi_2=0.0854$; $\pi_3=0.0933$; $\pi_4=0.1000$.
[c] $\pi_0=0.0500$; $\pi_1=0.0625$; $\pi_2=0.0750$; $\pi_3=0.0875$; $\pi_4=0.1000$.
[d] $\pi_0=0.0500$; $\pi_1=0.0531$; $\pi_2=0.0625$; $\pi_3=0.0781$; $\pi_4=0.1000$.

in advance of the current experiment because the powers of two tests can be evaluated from λ_a and λ_c, which can be calculated from the design of the current experiment, the target alternatives, and the realized values of the historical control which specify the range of the nuisance parameter π_0.

Note that, if we use exact conditional probabilities given Y and ΣX_i as *p*-values instead of the normal approximation, the test with the test statistic T_h automatically controls type I errors within a target significance level. Though the above argument is limited to a special case of binomial responses, it can be easily extended to other cases.

Rare Occurrence of Responses

Cases of the stable historical control discussed in the previous section rarely occur in toxicological experiments. We have seen such cases only in *in vitro* experiments or in short-term experiments. In most cases, there are more or less variabilities among historical controls and the concurrent control. Considering these variabilities as a prior distribution, Tarone (*2*) proposed a trend test based on the β-binomial model. Yanagawa and Hoel (*3*) proposed a set of trend tests based on the same line of thought as Tarone. In addition, the latter authors proposed exact test procedures that control the type I error when the asymptotic theory is not applicable.

To make the argument simple, let us assume a logistic response model on the rate parameter π_i for the binary response, that is,

$$\pi_i = \frac{\exp\{\gamma_0+\delta(d_i-d_0)\}}{1+\exp\{\gamma_0+\delta(d_i-d_0)\}}, \quad i = 0, 1, \ldots, a$$

$$\pi_{(j)} = \frac{\exp\{\gamma_j\}}{1+\exp\{\gamma_j\}}, \quad j = 1, 2, \ldots, b, \tag{18}$$

when the notation is the same as the one in the previous section. In the β-binomial model, the parameters $\pi_0, \pi_1, \cdots, \pi_{(b)}$, are assumed to be independent random variables with the density function

$$f(\pi) = \Gamma(\alpha+\beta)\pi^{\alpha-1}(1-\pi)^{\beta-1}/[\Gamma(\alpha)\Gamma(\beta)]. \tag{19}$$

If the two parameters α and β of Equation 19 are known, no problem arises in the proposed procedures. In real situations, however, these parameters are unknown and must be estimated from the historical control. Tamura and Young (*5*) pointed out that the estimation error seriously affects the type I error of Tarone's procedure. This sensitivity was also recognized by Yanagawa and Hoel (*3*). To overcome this defect, Krewski et al. (*4*) proposed a two-stage procedure that contains an optionally chosen parameter for the choice of the second-stage test. Though the result of the procedure is highly influenced by this parameter value, they do not give any reasonable method to determine it. Not doing so makes their procedure ambiguous.

In the same situations, Yanagawa and Hoel (*3*) suggested conservative use of the historical control. Yanagawa et al. (*9*) proposed two practical procedures in this line of thought and presented evidence for their effectiveness in some cases. Their idea is to make confidence intervals of two functions $\{\alpha+\beta, \alpha/(\alpha+\beta)\}$ of parameters and use a set of values of upper or lower confidence limit as the parameter values of the test statistic T_y defined below.

$$T_y = \sum_{i=0}^{a}(d_i-d_0)X_i - \frac{X+\alpha}{na+n+\alpha+\beta}\sum_{i=0}^{a}(d_i-d_0)n \tag{20}$$

where $X = \Sigma X_i$. Along this statistic, exact probabilities are accumulated, where "exact" means that conditional probabilities given X are calculated. By taking either the upper or the lower limit, four exact *p*-values are obtained for the test statistic T_y. The use of the maximum value of these values as the *p*-value for testing the hypothesis H_0: $\delta=0$ is Yanagawa et al.'s proposal (*9*). Let us denote this testing procedure by Ty.

In the test Ty, the historical control is used only to estimate α and β, and the estimated values of α and β are used as given constants. As a result, all the information contained in the historical control is included in the given condition, and so the resulting procedure adapts to our viewpoint. The conservative use of the historical control to keep type I errors within a target significance level is entirely the same idea as the one explained in the previous section. Therefore, the test Ty is recommended.

The problem is how to judge whether the test Ty is superior to the corresponding test, say Te, without the use of the historical control. It must be reasonable to assume that, in the test Te, the *p*-value is calculated by accumulating exact probabilities along the statistic T_e defined below.

$$T_e = \sum_{i=0}^{a}(d_i-d_0)X_i - \frac{X}{na+n}\sum_{i=0}^{a}(d_i-d_0)n. \tag{21}$$

The comparison of two tests can be carried out in the following manner. Define two sets S_y and S_e as

$$S_y = \{x \mid p_y(x) < \alpha, \; p_e(x) > \alpha\}$$
$$S_e = \{x \mid p_y(x) > \alpha, \; p_e(x) < \alpha\} \tag{22}$$

where $p_y(x)$ and $p_e(x)$ are *p*-values corresponding to T_y and T_e, respectively and α is the target significance level. If a target alternative H_1 is set, we compare exact probabilities $Pr\{S_y \mid H_1\}$ and $Pr\{S_e \mid H_1\}$. If $Pr\{S_e \mid H_1\}$ is greater than $Pr\{S_y \mid H_1\}$, then the test Ty should be used. The choice is possible in advance of the current experiment because the powers of the two tests can be evaluated from the design of the current experiment, the target alternatives, and the realized values of the historical control, which specify the range of the nuisance parameter π_0. Though we have not confirmed it, the method of Monte-Carlo simulation seems to be effective to evaluate probabilities.

Small Group Size

In toxicological experiments using large animals such as dogs or monkeys, the group size is as small as three or four. In such cases, any testing procedure of a hypothesis rarely yields significant results due to the lack of power. The incorporation of the historical control is highly attractive to increase the power of the test in this situation. This is the case C3 mentioned above.

In this case, we naturally assume that the observed response

is quantitative. Assume, as before, that the current experiment consists of $(a+1)$ groups of size n with dose d_i and observed responses are independent normal variables. Let the j^{th} response of the group A_i be X_{ij} and assume that the structural model of X_{ij} is the usual block-effect model such that

$$X_{ij} = \mu + \alpha_0 + \beta(d_i - d_0) + U_{ij} \quad (23)$$

where μ and β are unknown parameters, α_0 is distributed as $N(0, \Sigma_\alpha^2)$, and U_{ij} is distributed independently of α_0 as $N(0, \Sigma_\mu^2)$. Similarily, assume on jth response Y_{ij} of the ith historical control group the model

$$Y_{ij} = \mu + \alpha_i + V_{ij} \quad (24)$$

where μ and β are unknown parameters, α_i is distributed independently of α_0 as $N(0,\sigma_\alpha^2)$, and V_{ij} is distributed independently of α and U as $N(0,\sigma_u^2)$.

According to Margolin and Risko (*4*), when σ_α^2 and σ_u^2 are known, the maximum likelihood estimator of β is given by

$$\hat{\beta} = w\hat{\beta}_c + (1-w)\hat{\beta}_h \quad (25)$$

where $\hat{\beta}_c$ is the usual estimator of β based on the current experiment, $\hat{\beta}_h$ is the interblock estimator of β based on the historical control and the overall mean of X's, and the weight w is a monotone increasing function of $\sigma_\alpha^2/\sigma_\mu^2$, whose explicit form is shown in Margolin and Risko (*4*). If the historical control can be regarded as random quantities, then we can test the hypothesis $H_0:\beta=0$ by standardizing $\hat{\beta}$ with the square root of $Var(\hat{\beta})$, but that is not consistent with our viewpoint. $\hat{\beta}_h$ should be regarded as a given constant more or less deviated from the true value of β.

In addition, σ_α^2 and σ_u^2 are unknown in real situations. They are estimated from the historical control and within-group variances in the current experiment. This yields an estimation error on the weight w and causes an inflation or deflation of type I errors. We have to devise a conservative procedure to control the type I error within a target significance level. In principle, the same idea as the one in the previous section is available, that is, to use confidence limits as true values of w; but its realization has not been achieved up to now. Proposals of practical procedures are left for future studies.

Historical Control as a Reference

In chronic toxicity studies, hundreds of items are inspected during a long time interval within one experiment. This brings about many repetitions of statistical tests and causes an inflation of type I errors or frequent occurrences of false-positive results due to the multiplicity of tests. Encountering such errors, toxicologists do not usually accept the results of statistical data analysis unless the results are confirmed by toxicological and/or biological knowledge. When a toxicologist believes an observed statistical significance to be a realization of a type I error, he or she wants to validate this belief with evidence. In such circumstances, the historical control is used as evidence of the false positive of the statistical test. In fact, Matsumoto (*10*) found and reported many such cases through a survey of a volume of a journal. This is the case C4 mentioned in the initial section.

Our opinion on the use of the historical conrol in this case is rather negative because the variability among experiments is so big that the observed deviation of a treatment group from the concurrent conotrol group can be neglected almost always by using the historical control as the reference. This fact violates the rationality of the statistical reasoning. In this case, we recommend, in principle, the use of the distribution of *p*-values (*11*) to evaluate the inflation of the type I errors or to reduce many items to a few end points to avoid multiplicities (*12*), though the construction of practical procedures is not easy.

Concluding Remarks

Two points are emphasized in this paper: one is that the historical control should be regarded as a given condition and the other is that it should be used conservatively. We recommend the incorporation of historical controls only when it is advantageous under such a conditional evaluation of the performance; even then it should be used conservatively.

In this paper, we considered only simple situations and simple procedures. In such cases, the choice of whether to adopt the incorporation is not difficult because the performance of the two procedures can be, at least approximately, evaluated and compared based on the design of the current experiment, the target alternatives, and the realized values of the historical control, which are obtained in advance. The application of this viewpoint seems easy for more complicated situations if we concentrate our attention on simple procedures. In real situations, however, there is a possibility that a more complex procedure adapts to our viewpoint better than such simple procedures. One example of this is shown in Hayashi et al. (*1*). The evaluation of such complex procedures is left for future investigations.

REFERENCES

1. Hayashi, M., Yoshimura, I., Sofuni, T., and Ishidate, M., Jr. A procedure for data analysis of the rodent micronucleus test involving a historical control. Environ. Mol. Mutagen. 13: 347-356 (1989).
2. Tarone, R. E. The use of historical control information in testing for a trend in proportions. Biometrics 38: 215-220 (1982).
3. Yanagawa, T., and Hoel, D. G. Use of historical controls for animal experiments. Environ. Health Perspect. 63: 217-224 (1985).
4. Margolin, B. H., and Risko, K. J. The use of historical data in laboratory studies. Proc. Int. Bio. Conf. 21-30 (1984).
5. Tamura, N., and Young, S. S. The incorporation of historical control information in tests of proportions: simulation study of Tarone's procedure. Biometrics 42: 343-349 (1986).
6. Hoel, D. G., and Yanagawa, T. Incorporating historical controls in testing for a trend in proportions. J. Am. Stat. Assoc. 81: 1095-1099 (1986).
7. Krewski, D., Smythe, R. T., and Colin, D. Tests for trend in binomial proportions with historical controls: a proposed two-stage procedure. In: Biostatistics (J. B. MacNeill and G. J. Umphrey, Eds.), Reidel Publishing Co., Dordreht , The Netherlands, 1987, pp. 61-69.
8. Yanagawa, T., Hoel, D. G., and Books, G. T. A conservative use of historical data for a trend test in proportions. J. Jn. Stat. Soc. 19: 83-94 (1989).
9. Cox, D. R. The regression analysis of binary sequences. Jr. R. Stat. Soc. B 20: 215-231 (1958).
10. Matsumoto, K. Some problems on the use of historical controls (in Japanese). Presented at the Biostatistics Symposium for the Analysis of Toxilogical Test Data, Tokyo, November 14-15, 1990. (unpublished)
11. Selwyn, M. R. Dual controls, *p*-value plots, and the multiple testing issue in carcinogenicity studies. Environ. Health Perspect. 82: 337-344 (1989).
12. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. The analysis of multiple endpoints in clinical trials. Biometrics 43: 487-498 (1987).